Genetic diversity of 100+ Aspergillus species: The aspMine analysis resource

Vesth, Tammi Camilla; Rasmussen, Jane Lind Nybo; Theobald, Sebastian; Kjærbølling, Inge; Frisvad, Jens Christian; Nielsen, Kristian Fog; Lyhne, Ellen Kirstine; Kogle, Martin Engelhard; Kuo, Alan; Riley, Robert

Total number of authors:
16

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Genetic diversity of 100+ Aspergillus species

The aspMine analysis resource

Tammi C. Vesth* (1), Jane L. Nybo (1), Sebastian Theobald (1), Inge Kjærbølling (1), Jens C. Frisvad (1), Kristian F. Nielsen (1), Ellen K. Lyhne (1), Martin E. Kogle (1), Alan Kuo (3), Robert Riley (3), R.P. de Vries (4), Igor V. Grigoriev (3), Uffe H. Mortensen (1), Bernard Henrissat (5), Scott E. Baker (2), Mikael R. Andersen (1)

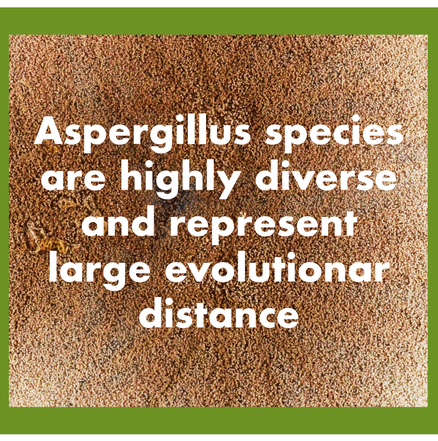
1) Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark
2) Joint Bioenergy Institute, Berkeley, CA, USA
3) Joint Genome Institute, Walnut Creek, CA, USA
4) Fungal Physiology, Westerdijk Fungal Biodiversity Institute - KNAW Fungal Biodiversity Centre, Utrecht, the Netherlands

Technical University of Denmark

Curious about the Aspmine? Ask me!

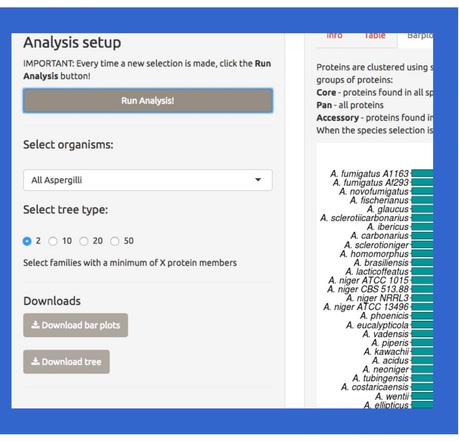
tcve@bio.dtu.dk

Background



Interactive analysis apps

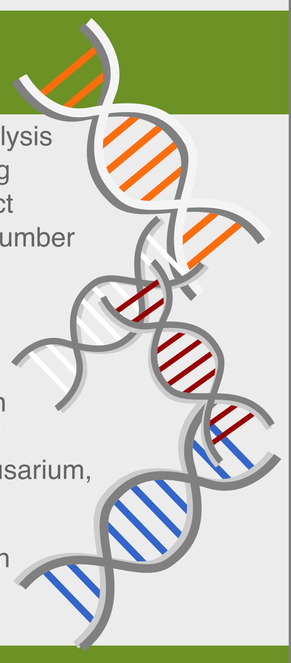
- The aspmine is a webpage
- Contains documentation of analysis
- Holds link to interactive analysis apps
- Apps allow the user to explore the analysis data
- Figures and tables are available for download
- Analysis in the apps can be customized by selecting organisms of interest and cutoffs or subsets of data



Fungi

- Producers of chemically and medically relevant compunds
- Large natural diversity and high production of bioactive compounds
- Well studied production organisms
- Can be genetically optimized for production of cheeper and environmentally friendly compounds

Sequencing



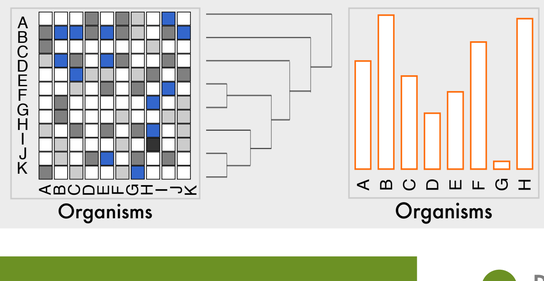
Genome sequencing and analysis can elucidate many interesting genetic features. In this project we aim to sequence a large number of Aspergillus species

- Initiative to sequence > 300 species
- DTU IBT Culture Collection > 35,000 fungal cultures of Penicillium, Aspergillus, Fusarium, Alternaria and Tricoderma
- Data represents 200 million years of evolution

Comparative Genomics

Comparative genomics can be used to investigate a number of aspects of genetic diversity. Here we focus on evolutionary development of Aspergillus species and diversity of secondary metabolism

- Genomes/annotation measures and quality
- Genetic diversity of Aspergilli
- Families of proteins with shared functions
- Genes/proteins specific to single species
- Shared protein families within groups of species
- Horizontal gene transfers across large phylogenetic distances
- Families of secondary metabolism gene clusters responsible for similar compounds



Analysis setup

Analysis database - the workhorse

- Relational databases are ideal for combining different data types as well as conditional selction of data
- Raw data is obtained from JGI and is stored in a costume designed relational database (MySQL)
- Data include functional annotation and protein sequences
- All proteins are compared across the growing dataset (BLAST)
- BLAST alignement scores are stored in the database

Raw information
- Taxonomical section
- Protein sequences
- Organism name and ID

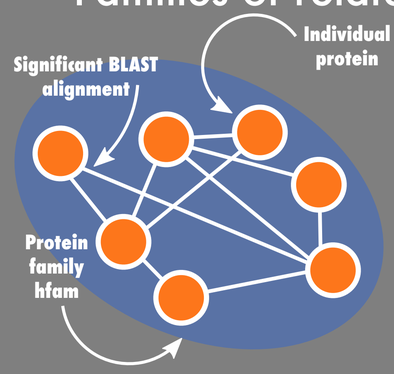
Functional annotation
- SMURF - secondary metabolism
- InterPro - domains
- KOG - kategories
- KEGG - pathways

BLAST alignment data
- Query/hit coverage
- Alignment coverage
- Percent identity
- Reciprocal hit flag

Selecting data conditionally

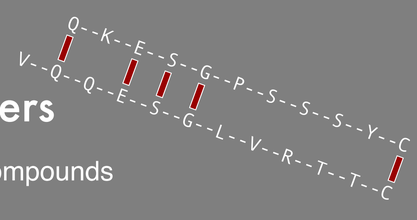
- MySQL is ideal for extracting data which fulfills specific criteria
- Examples of conditonal selection of data in the analysis database
  - Genes found only in a subset of species
  - Specific functional annotation which i always found twice in Aspergilli
  - Genes containing signal peptides and found in a specific set of species

Families of related proteins - aspmine hfams



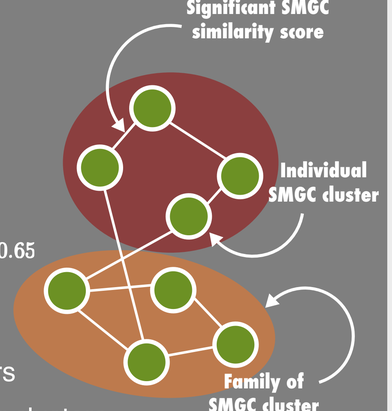
- Assumption: similar protein sequences imply similar function
- Sequence similarity can be defined by alignments
- Alignements are determined using BLAST
- Alignement coverage of hit plus query must be over 130%
- Identical resedues in alignment must be over 50%
- Proteins are connected using single linkage, a protein need only be connected to one other memeber of the cluster.

Families of related SMGCs - Secondary Metabolism Gene Clusters



- Assumption: SMGCs with similar genes will create similar compounds
- Functionally similar genes can be identified using BLAST
- Enzymes that initiate a secondary metabolite (backbones) are most important in the definition of that metabolite
- Tailoring enzymes also hold information about the metabolite but to a lesser extend than backbone enzymes
- SMGCs must share a significant fraction of backbone and tailoring enzyme activities to create similar compounds

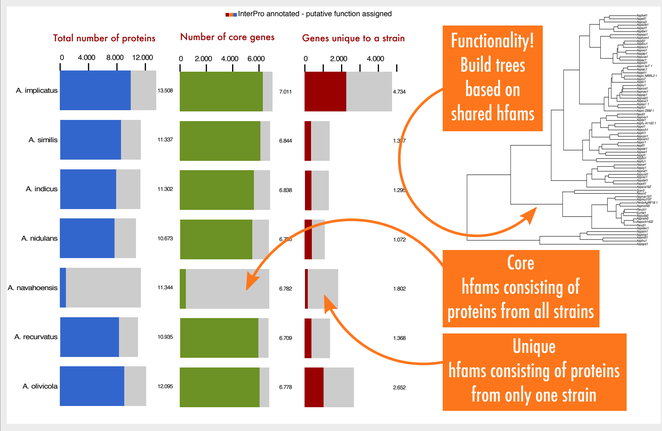
$$\textbf{Pairwise SMGC similarity score} = pident_{tailoring} * \frac{count_{tailoring}}{total_{tailoring}} * 0.35 + pident_{backbone} * \frac{count_{backbone}}{total_{backbone}} * 0.65$$



- SMGCs are predicted using the SMURF algorithm
- A costum score is used to calculate the similarity between clusters
- Two rounds of random walk clustering connectes the most similar clusters and creates a network which illustrates the interconnectedness of SMGCs

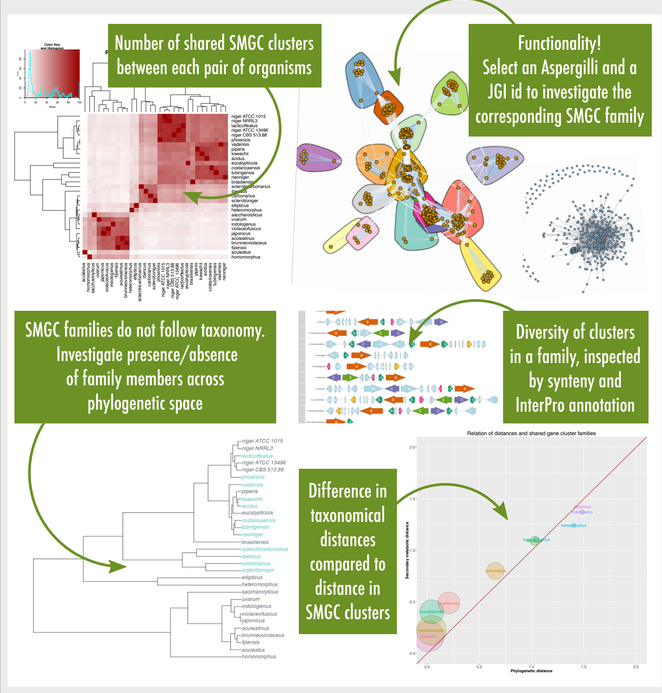
Genetic Diversity

- Analysis apps are available for analysis of genetic diversity through the construction of protein families - aspmine hfams
- DNA/protein sequence comparisons are essenitial to comparative genomcis
- Proteins with similar sequences form clusters of functinally related proteins - protein families
- Closely related strains share more families
- Many families are strain specific!



Secondary Metabolism

- Analysis apps are available for analysis of SMGC families and cluster variation across species of Aspergilli
- SMGC families can be quried using protein FASTA sequences or JGI protein identifiers
- Families can be inspected by synteny plots illustrating conserved functions and organization
- SMGC clusters do not follow the standard taxonomy and their presense across phylogeny can be explored



- Rshiny is a R package for interactive web apps
- Rshiny can be hosted on www.shinyapps.io
- Graphical interphases for R data and graphics
- Cheap hosting of interactive web-applications
- Analysis with customization and documentation
  - Hosted with www.shinyapps.io ($440 USD/year)
  - Unlimited apps, 500 active hours

aspmine.wordpress.com

Documentation

- Documentation of data analysis is often neglected
- Publication of thorough data methods is insufficient
- A webpage is a good place to document analysis
- The Asp Mine offers online documentation
- Access to analysis data and descriptions of methods